An approach to estimate degree completion using drop-out rates

Stijn Luca[a,*], Marc Verdyck[b], Marc Coppens[c]

[a] *Department of Electrical Engineering (ESAT), KU Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium*
[b] *Thomas More Kempen, Kleinhoefstraat 4 2440 Geel, Belgium*
[c] *Department of Mathematics, KU Leuven, Celestijnenlaan 200B, B-3001 Leuven, Belgium*

Abstract

A completion rate of an institution I of year x is defined as the proportion of starting students during year x that succeed in graduating at I at some point in the future. In this paper, a new method is proposed to estimate such completion rate. This indicator is entirely based on the population of drop-out students during one academic year x at the institution I. The proposed method is not based on a cohort of students so it allows an estimation of the current school effectiveness without substantial delay. Furthermore a statistical framework is presented in which completion rates can be studied. The proposed methodology results in a statistical estimator with a bias that stays small under appropriate assumptions.

Keywords: completion rate, graduation rate, binomial distribution, school effectiveness, student success

1. Introduction

Student success plays a crucial role in student careers, the accountability of educational instutions and the development of our society. Completion rates have a long history of being used as indicators of student success. In this paper a new measure for completion rate is studied. A completion rate of an institution I of year x is defined as the proportion of starting students during year x that succeed in graduating at I at some point in the future.

In post-secondary institutions, such a completion rate plays a more crucial role than an on-time graduation rate. Here it is more common that students choose to extend their study period by one or more semesters, e.g. in order to be able to combine studying with a part-time job.

Ideally an unbiased estimator of completion rate can be obtained in a cohort study where students are individually tracked through time. The duration of such cohort implies a substantial delay on the calculation of the completion rate. Moreover the obtained measure does not reflect the current state of school effectiveness. On the contrary, it is rather a result of the education process during the complete cohort period.

This paper aims for two main contributions. On one hand a new measure for completion rate of an institution I is developed. In contrast to cohort studies, substantial delay is avoided and the most up-to-date information that is available is used. For this purpose a measure is developed that is completely based on the population of drop-out students during a specific academic year x, e.g. the current year.

On the other hand the performance of this measure is studied as a statistical estimator for completion rate that does not depend on study duration. The estimator is investigated in a statistical framework to study under which assumptions statistical inferences can be made.

This paper discusses completion rate in an educational context. However completion rates are also found in other disciplines. In computer science for instance a completion rate can be viewed as the number of complete files that are succesfully transferred from one server to another. Therefore it should be noted that what follows can be adapted for its use in other disciplines.

2. Completion rates in education, a review

Very related to completion rates are on-time graduation rates that depend on study-duration. In the United States all institutions of higher education are required by law to publish completion rates. For this reason there is a rich literature on an on-time graduation rate as defined in the No Child Left Behind Act (NLCB, 2002). However in using such graduation rates as a measure for school effectiveness caution is needed. As suggested in Astin (2005), on-time graduation rates and more generally completion rates cannot be adequately adressed without considering the kinds of students who initially enroll.

Nonetheless, differences among the measures that are published has lead to much debate in the United States over the correct rate to use to meet the demands of the NCLB act (Swanson & Chaplin, 2003). The two main difficulties in the calculation of completion rates are a lack of comprehensive sources for data on completion and a lack of consensus on the conceptual and technical definitions related to completion rates. An extensive overview of this matter is given in Hauser & Koenig (2011).

*Corresponding author
Email address: stijn.luca@kuleuven.be (Stijn Luca)

Preprint submitted to Studies in Educational Evaluation on 15/04/2013

Ideally completion rates can be calculated in a cohort study. In such studies a set of starting students is followed during their study careers and completion rates are calculated after graduation (see e.g. Ensminger & Slucarcick (1992); Boden (2011-2012)). In practice however this leads to several problems. A cohort study implies the need of longitudinal databases that track individual students through time. The availability or collection of such a dataset is far from obvious (Hauser & Koenig, 2011).

Moreover the duration of the cohort immediately implies a substantial delay on the calculation of a completion rate. This delay can be avoided by the use of an estimator that is based on available data through current and historical records. Such estimations do not require the tracking of individual student over time. However, they can only be viewed as proxies for a true cohort indicator.

In literature many estimator for on-time graduation rates are found. Comprehensive overviews can be found in Miao & Haney (2004); Seastrom et al. (2006); Swanson & Chaplin (2003). They range from very simple estimators to more complicated estimators wich also require more complex demands on data collection systems.

The simple on-time graduaton rate (SGR) for instance is defined as the ratio of the number of students G_x that graduate during year x and the total number n_{x-T} of enrollments at institution I in year $x-T$ where T denotes the duration of the study-program under consideration:

$$SGR = \frac{G_x}{n_{x-T}} \tag{1}$$

This simple estimator will likely differ from the true cohort rate because students move in and out of the institution I during the period of T years. Alternatives proposed by Greene & Winters (2002) and Haney (2001) try to reduce this bias by adjusting the number of enrollments in the denominator.

The 'cohort graduation rate' and 'exclusion-adjusted cohort graduation indicator' (Seastrom et al., 2006), the latter being proposed by the National Institute of Statiscal Sciences, are examples of cohort graduation rates requiring detailed data on individual students over time.

Swanson & Chaplin (2003) developped the Cumulative Promotion Index for high school graduation rate. This indicator estimates on-time graduation rate as a probability that a starting student will graduate on time. The estimator is based on a so-called synthetic cohort consisting of shortened time periods and only requires data from two school years.

In this paper a new estimator for completion rate is proposed that is completely based on the current population of drop-out students. In contrast to estimators for an on-time graduation rate, commonly found in literature, the indicator does acount for students that graduate in a period that is longer than the standard number of years. It is therefore more suitable for post-secondary institutions.

The 'NCES high school completion rate' is another indicator that is based on drop-out counts. However this rate requires drop-out counts in each level of the study-program over the last T-years. Such drop-out statistics are often not available (Seastrom et al., 2006).

3. From drop-out rates to completion rates

In this section a method is developed to estimate the completion rate of an institution I. This rate is defined as the proportion of starting students, that are recruited by an institution I during an academic year x and will graduate at some point in the future.

In particular the goal in this section is to develop an estimation in a way that:

1. The most up-to-date information that is available is used. This is in contrast with existing estimator that are mainly based on historical records.
2. The method is not based on a cohort study. Therefore, information is obtained without substantial delay.
3. The estimation is independent of the study duration and therefore very useful in post-secondary institutions.

3.1. Methodology

Up-to-date information can be found by retrieving information from the set of students that drop out during academic year x. The number of these drop-out students can be calculated after the end of the enrollment period of academic year $x+1$ by registering those students enrolled during year x but did not reregister the subsequent year. In this way a delay of maximal one year is obtained in contrast to cohort studies where several years are needed to obtain an estimation.

This set can be subdivided into subsets according to the academic year $x-i$ (a previous academic year where i denotes a natural number) during which a student started his study at the institution I. Figure 1 illustrates such partitioning of the population of drop-out students.

When the school is performing well, the set of drop-out students that started during a year $x-i$ should be small. One also expects that the size of the subsets decrease with i and will be negligible from some index $i = T$. Generally it will suffice to apply the partition up till $5-6$ subsets (i.e. $T = 3-4$). Denote s as the sum of the ratios obtained by dividing the sizes of these subsets by the number of starting students during the academic years $x-i$. The complement of this sum $1-s$ is the new proposed estimation of the completion rate of I.

3.2. Example

For illustrative purposes an example of the use of the formula is presented based on data of Thomas More University College (UC). The drop-out ratio's presented in table 1 were calculated until the academic year $2010-2011$. The table presents the drop-out ratio's according to the

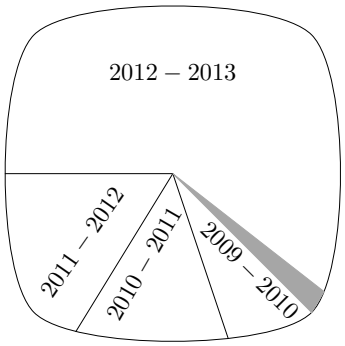

Figure 1: Schematic illustration of the partition of the population of drop-out students. Academic year x is taken as the academic year, $x = 2012 - 2013$. For $T = 3$ the population of drop out-students is partitioned into 5 subsets according to the starting years. Students that drop-out at year x and have started more than 4 years ago correspond to the shaded area and are expected to be small.

Drop-out ratio (%)		Number of academic years from enrollment to drop-out				
		1	2	3	4	5
Starting year	2006-2007	**29.48**	**8.71**	**3.92**	**1.06**	**0.37**
	2007-2008	28.23	6.70	2.98	1.16	
	2008-2009	25.18	8.88	2.91		
	2009-2010	26.93	7.65			
	2010-2011	25.17				

Table 1: Table of proportions of drop-out students at Thomas More UC according to starting year and graduating year.

starting year and the number of years from enrollment to drop-out. Most standard study-programmes at Thomas More UC involve a full time study period of length 3 years. The sum of the ratio's in the colored boxes of table 1 correspond to the sum s, described in section 3.1. Application of the proposed method leads to an estimation of a completion rate of the academic year $2010 - 2011$ given by 62.74%.

The horizontal bar of bold numbers corresponds to a cohort that could have been followed from the start of the academic year $2006 - 2007$ leading to a completion rate of 56.46%. This true cohort rate of $2006 - 2007$ could have been estimated using the new method proposed above, based on the population of drop-out students of $2006 - 2007$. Research of the historical database of the UC led to a completion rate of 55.59%. Merely a bias of 0.87%.

The real power of the method lies in the fact that most students drop out in their first year. Because this ratio is the biggest in the sum s, the estimation emphasizes the most up-to-date information that is available. This is a major difference that distinguishes the new proposed methodology from other indicators found in literature.

To illustrate this further, consider an estimation process using graduation rates instead of drop-out rates. Table 2 shows completion data which was only made available by the administration until academic year $2009 - 2010$. The graduation ratio's are presented according to the starting year and the number of years needed to graduate. The completion rate of $2009 - 2010$ can be estimated as the sum of colored diagonal elements leading to 55.7%. However this sum emphasizes the proportion 36.40% that is rather due to historical events than recent events. Also a true cohort estimation (sum of bold numbers) emphasis a proportion during the academic year $2007 - 2008$.

Time to graduation ratio (%)		Number of academic years from enrollment to graduation				
		1	2	3	4	5
Starting year	2005-2006	**3.01**	**1.44**	**36.34**	**12.97**	**2.33**
	2006-2007	0.56	1.76	39.13	12.86	
	2007-2008	3.26	3.38	36.40		
	2008-2009	0.68	3.14			
	2009-2010	0.97				

Table 2: Table of proportions of graduated students at Thomas More UC according to starting year and graduating year.

3.3. Discussion: assumptions and limitations

The example in the previous section showed a small bias between a true cohort estimation of completion rate and an estimation based on the new methodology. In what follows the underlying assumptions are discussed that have to be met in order to obtain a small bias. In the next section these assumptions will be re-encountered when placing our methodology in a statistical framework.

Firstly, the drop-out rate among students during year x that subscribed during year $x - i$ has to approximate the drop out rate of current first-year students who survive to their ith year and this for each integer i between 1 and T. Generally, these assumptions will be met if circumstances related to curriculum and characteristics of newly entering students did not change drastically during the past T years and at the same time will not undergo major changes during the next T years.

Secondly so-called stop-outs could induce bias if they manage to complete their degree. Stop-out students drop out before completing a degree and return during the period under study. This limitation however is also encountered in estimators of on-time graduation rate like that of Greene & Winters (2002) or Haney (2001). If the necessary information is available the new estimator can be corrected for this bias by subtracting the numbers of stop-outs that complete their degree from the denominator for the year in which they dropped out.

A severe bias can be introduced when these assumptions are not met. However, information is obtained without substantial delay, which is not the case for cohort studies. Moreover the method is based on the current population of drop-out students which imply, as is illustrated in the previous example, that the most up-to-date information that is available is used. This can be of high value for

policy makers who wish to obtain a measurement based on current information rather than on historical records.

Although the new estimator proposed offers new advantages one can encounter problems due to data limitations. In order to be able to implement the method, the institution needs to know for each currently enrolled student:

1. the year the student first enrolled,
2. whether the student is a first-time student, a transfer student, or a returning stop out.

Moreover it can be interesting in keeping track whether students initially enrolled as a full-time or a part-time student. In this way, it is possible to perform separate studies on these two groups.

4. Formal framework

In this section the new estimator of completion rate is studied in a statistical framework. The assumptions mentioned in section 3.3 will be retrieved as mathematical assumptions in section 4.2. This will make it possible to test the bias for statistical significance in section 4.3. The formal framework starts with presenting a set of statistical definitions which allow the definition of degree completion in general terms.

4.1. Definitions

Given an academic year x and an educational institution I, the purpose is to gain information of the completion rate, i.e. the proportion of students that start their study at I during year x and obtain their degree at that same institution at some point in the future.

In the following definitions $\mathcal{A}$ denotes the population from which I can recruit.

Definition 1.

- *A* starting student *of academic year x is a student from $\mathcal{A}$ that is enrolled for the very first time at the institution I during year x in a regular enrollment period.*
- *An* effective student *is a starting student from $\mathcal{A}$ that succeeds in graduating at some point in the future.*

Statistically each starting student of academic year x has a probability of being an effective student. Formally the following definition is stated.

Definition 2. *The* completion probability p_x *with respect to academic year x is defined as the probability that a starting student from $\mathcal{A}$, that is recruited during academic year x, is an effective student.*

The number of starting students G_x of an academic year x that are effective students is a random variable that is distributed according to a binomial distribution:

$$G_x \sim B(n_x, p_x) \tag{2}$$

where n_x denotes the number of starting students and $p_x \in [0, 1]$ the completion probability of academic year x.

Complementary to the random variable G_x is the number of *drop-out students* that leave the institution I before graduating.

Definition 3.

- *A drop-out student is a starting student from $\mathcal{A}$ that is not effective.*
- *The* drop-out probability r_x *with respect to academic year x is defined as the probability that a starting student from $\mathcal{A}$, that is recruited during academic year x, is not effective.*

The amount of starting students of academic year x that are drop-out students is denoted as D_x. Also D_x is a random variable that is distributed according to a binomial distribution:

$$D_x \sim B(n_x, r_x), \tag{3}$$

where n_x denotes the number of starting students and $r_x \in [0, 1]$ the drop-out probability of academic year x. Obviously

$$D_x + G_x = n_x$$

and $r_x = 1 - p_x$.

The parameters p_x and r_x are theoretical quantities. The proces of estimating these quantities using sample data is called *statistical inference* (Cox, 2006). An estimator of p_x (resp. r_x) can be seen as a random variable $\widehat{P}_x$ (resp. $\widehat{R}_x$) whose outcome depends on random sample data. These estimators are called unbiased when:

$$E(\widehat{P}_x) = p_x \quad \text{and} \quad E(\widehat{R}_x) = r_x$$

where E denotes the expectation operator. A natural and common estimator of p_x (resp. r_x) is given by the completion rate (resp. drop-out rate). These rates correspond to the so-called maximum likelihood estimators (MLE).

Definition 4.

- *The* completion rate *with respect to an academic year x is the ratio of the number of effective students G_x over the number of starting students n_x:*

$$\overline{G}_x := \frac{G_x}{n_x}$$

- *The* drop-out rate *with respect to an academic year x is the ratio of the number of drop-out students over the number of new students n_x:*

$$\overline{D}_x := \frac{D_x}{n_x}$$

Obviously $E(G_x) = n_x p_x$ and $E(D_x) = n_x r_x$ so that the MLE's are unbiased meaning:

$$E(\overline{G}_x) = p_x \quad \text{and} \quad E(\overline{D}_x) = r_x$$

4.2. Estimating completion rate

Without a crystal ball, direct calculation without delay of the estimators in definition 4 is not possible. All estimators discussed in section 2 are attempts to construct estimators of p_x or r_x. Also in this paper, a new estimator of r_x was informally introduced in section 3. In this section a formal description is given of the methodology introduced in section 3.

Consider first the case where one would obtain an estimation of a drop-out probability r_x using a cohort study. This means that a group of students would be followed during some period of T years. For the estimation to be unbiased the number of years T has to be chosen large enough to allow counting all drop-out students starting at academic year x.

At the end of the cohort study, the number of drop-out students can be thought of as the realisation of a sum of random variables:

$$D_x = D_x^0 + D_x^1 + D_x^2 + \cdots + D_x^T \tag{4}$$

where D_x^i is the random variable associated with the starting students of academic year x that will leave I during academic year $x + i$.

The random variables D_x^i $(0 \leq i \leq T)$ are distributed according to:

$$D_x^i \sim B(n_x, r_x^i)$$

where r_x^i denotes the drop-out probability of a starting student who will leave I during academic year $x + i$. Obviously:

$$r_x = r_x^0 + r_x^1 + r_x^2 + \cdots + r_x^T \tag{5}$$

In a cohort study over a period of T years, the drop-out probability r_x would therefore be estimated using:

$$\overline{D}_x = \sum_{i=0}^{T} \overline{D}_x^i$$

A cohort study proceeds by calculating each $\overline{D}_x^i$ by tracking a set of starting student through time over the period $[0, T]$. When x is the current academic year, this estimator implies a substantial delay. Instead an estimator for r_x was described in section 3 using the most up-to-date information available. Let us proceed with a statistical description of this methodology.

For this purpose consider the set S of students at the institution I that abandon their studies during academic year x (during some regular enrollment period). Consider for each student in S the acadamic year he or she was first registered. Denote this year as $x - i$, $0 \leq i \leq T$. In this way the set S of drop-out students can be divided into subsets S_i according to their starting year in the educational institution I.

One can now propose the following estimator for r_x:

$$\widehat{R}_x = \overline{D}_x^0 + \overline{D}_{x-1}^1 + \overline{D}_{x-2}^2 + \cdots + \overline{D}_{x-T}^T \tag{6}$$

The sum of the ratios of the number of students in S_i divided by the number of starting students during academic year $x - i$ is a realisation of $\widehat{R}_x$.

This estimator meets the requirements of a formal description of the measure for completion rate described in section 3. The colored diagonal elements in table 1 denote realisations of $\overline{D}_{x-i}^i$ for $0 \leq i \leq 4$ with $T = 4$. In practice one chooses $T \in \mathbb{N}$ such that $\overline{D}_{x-T}^T$ is negligible.

In contrast to $\overline{D}_x$ the estimator $\widehat{R}_x$ can be statistically biased. From (5) and the linearity of the expectation operator, it follows:

$$E(\widehat{R}_x) = \sum_{i=0}^{T} r_{x-i}^i \qquad \text{and} \qquad E(\widehat{R}_x) - r_x = \sum_{i=0}^{T} (r_{x-i}^i - r_x^i) \tag{7}$$

The latter defines the bias of $\widehat{R}_x$ that remains small as long as $r_{x-i}^i \approx r_x^i$. These assumptions were already described in section 3.3 and are re-encountered here in a mathematical formula. Furthermore the variance of the estimator $\widehat{R}_x$ is given by:

$$\mathrm{Var}(\widehat{R}_x) = \sum_{i=0}^{T} \frac{r_{x-i}^i (1 - r_{x-i}^i)}{n_{x-i}} \tag{8}$$

Formula (7) implies a cumulative bias in which all biases $r_{x-i}^i - r_x^i$ are summed. In practice this can lead to a substantial bias of the estimator. In the following example we show results obtained from a study performed at Thomas More UC to illustrate the bias calculated in (7).

Example 1. Taking academic year 2006-2007 as year x, one can compare the results of a cohort study along a period of 5 years up to the academic year (2012–2013) with the results obtained by using the estimator (6). We perform a comparative study for two populations, one that counts a relatively small number of students and one containing a larger number of students.

Table 3 considers the students at Thomas More UC following a bachelor in Agro- and biotechnology. During the past 5 years, the bachelor degree in Agro- and biotechnology counted an average of 176 starting students.

Table 4 presents data corresponding to all students at Thomas More UC. The campus counts an average of 2317 starting students each year. Note that this table contains the numbers behind the ratios in the colored boxes of table 1. The example of section 3 is reviewed starting from formula (6).

Using the notation introduced in the sections above, the completion rate in a cohort study starting at year x (2006-2007) and ending in academic year 2011-2012 would be estimated as:

$$\overline{D}_x = \sum_{i=0}^{4} \frac{D_x^i}{n_x} \tag{9}$$

Bachelor Agro - and biotechnology				
i	D_x^i	n_x	D_{x-i}^i	n_{x-i}
0	51	178	51	178
1	21	178	20	163
2	16	178	11	189
3	5	178	5	224
4	0	178	0	85

Table 3: Number of drop-outs and starting students for the Bachelor in Agro - and biotechnology with reference year $x = 2006 - 2007$

Thomas More UC				
i	D_x^i	n_x	D_{x-i}^i	n_{x-i}
0	640	2171	640	2171
1	189	2171	218	2351
2	85	2171	77	2230
3	23	2171	44	2226
4	8	2171	5	2155

Table 4: Number of drop-outs and starting students for Thomas More UC with reference year $x = 2006 - 2007$

whereas using the estimator in formula (6) one would calculate:

$$\hat{R}_x = \sum_{i=0}^{4} \frac{D_{x-i}^i}{n_{x-i}} \tag{10}$$

where T is chosen as 4. The realisation of the difference $\hat{R}_x - \overline{D}_x$ leads to an estimation of the bias (7).

For the bachelor degree this leads to an estimation of 3.27%. For Thomas More UC there is only a bias of -0.87% (whose absolute value was already found in section 3.2). This smaller bias can indicate a better approximation or could be caused by chance.

4.3. Testing for significance of the bias

In a study as presented in example 1 one could be interested in a comparison of the two biases that we calculated, keeping in mind that the smaller bias on the estimate of the drop-out probability of Thomas More UC could be due to chance rather than to a better approximation.

In particular one wonders whether the difference between the estimators (9) and (10) leads to a bias that is significantly different from zero based on the underlying binomial distributions.

Formally a H_0–hypothese is tested that reads:

$$H_0 : \sum_{i=0}^{T} r_x^i = \sum_{i=0}^{T} r_{x-i}^i$$

To test such hypothesis, we will rely on normal approximations of the binomial distributions that underlie our method.

It is well known that the proportions $\overline{D}_x^i$ (6) can be approximated by a normal distribution with the same expectation and variance as $\overline{D}_x^i$:

$$\overline{D}_x^i \sim N(r_x^i, \frac{r_x^i(1-r_x^i)}{n_x})$$

Therefore, the estimator used in a cohort is approximately distributed according to:

$$\overline{D}_x = \sum_{i=0}^{T} \overline{D}_x^i \sim N\left(\sum_{i=0}^{T} r_x^i, \sum_{i=0}^{T} \frac{r_x^i(1-r_x^i)}{n_x}\right)$$

Analogous $\widehat{R}_x$ is approximately normally distributed with expectation and variance as in (7) and (8):

$$\widehat{R}_x \sim N\left(\sum_{i=0}^{T} r_{x-i}^i, \sum_{i=0}^{T} \frac{r_{x-i}^i(1-r_{x-i}^i)}{n_{x-i}}\right)$$

A general rule gives that these approximations can be applied under the conditions:

$$n_x \cdot \min\{r_x^i, 1-r_x^i\} \geq 5 \qquad \text{and} \qquad n_x \geq 30 \tag{11}$$

Under the H_0-hypothese, one finds:

$$\widehat{R}_x - \overline{D}_x \sim N\left(0, \sum_{i=0}^{T} \frac{r_{x-i}^i(1-r_{x-i}^i)}{n_{x-i}} + \frac{1}{n_x}\sum_{i=0}^{T} r_x^i(1-r_x^i)\right) \tag{12}$$

Based on this normal approximation, the H_0-hypothesis is rejected with significance level α when the empirical measure of the statistic:

$$\frac{\widehat{R}_x - \overline{D}_x}{\sqrt{\sum_{i=0}^{T} \frac{r_{x-i}^i(1-r_{x-i}^i)}{n_{x-i}} + \frac{1}{n_x}\sum_{i=0}^{T} r_x^i(1-r_x^i)}} \tag{13}$$

is outside the interval $[-z_{\frac{\alpha}{2}}, z_{\frac{\alpha}{2}}]$, with $z_{\frac{\alpha}{2}}$ the $(1-\frac{\alpha}{2})$–quantile of the standard normal distribution.

It is commonly known that the normal approximation of a binomial distribution $B(n,p)$ fails for small n, see for instance Sauro & Lewis (2005). The problem in this application is even more delicate because one encounters a sum of proportions. To verify whether the reliability of the approximation is satisfactory for practical purposes, one can use simulation.

Example 2. Based on the datasets of example 1 two simulations of the statistic in (13) are performed. To this purpose one mimics the binomial distributions $B(n_x, r_x^i)$ and $B(n_{x-i}, r_{x-i}^i)$ underlying this test statistic.

For a simulation of drop outs from the bachelor Agro- and Biotechnology, parameters of the binomial distribution are estimated using data in table 3. For each mimic of the underlying binomial distributions an estimation of the bias can be calculated from (13). The distribution of these calculated biases should approximately be normal as stated in (12). To this end it is typically to compare the quantiles obtained from this simulation to the theoretical quantiles extracted from a normal distribution. In figure

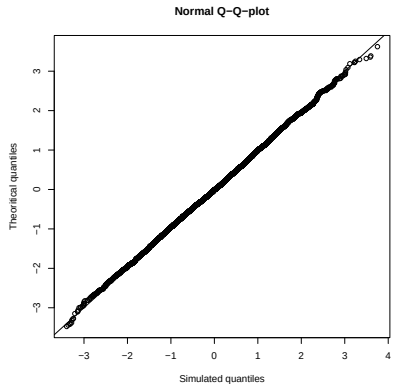

Figure 2: Quantile plot of simulated quantiles versus normal quantiles corresponding to the data collected from the Bachelor Agro- and Biotechnology.

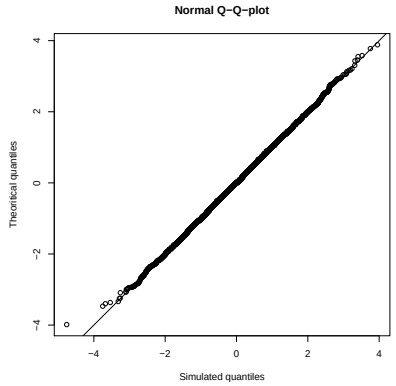

Figure 3: Quantile plot of simulated quantiles versus normal quantiles corresponding to the data collected from Thomas More UC.

2 a plot is shown of simulated versus theoretical quantiles (so called Q-Q plots, quantile-quantile plots). The simulation consists of 1000 quantiles. One expects that the graphed data follows the diagonal shown in the plot. The resulting plot shows therefore that the approximations is satisfactory for this example.

Figure 3 shows a Q-Q plot analogously obtained from simulation based on data of table 4. Also this plot is satisfactory.

From these results one is able to compare biases obtained in both examples. The bias of 3.27% for the bacholor degree leads to a realisation of the test statistis (13) given by 0.486. The bias of -0.87% for Thomas More UC leads to a test statistic of -0.492. Both biases are therefore comparable. Their absolute sizes are not big enough to reject our H_0–hypothese, keeping in mind that the 0.975-quantile of the standard normal distribution is approximately 1.96.

5. Conclusion

In this article we focused on developing a measure for completion rate. Many indicators for this purpose are in some way based on a cohort. In a prospective cohort study, the cohort is assessed at the beginning of the study and followed into the future. In this way a result is only possible after a substantial period in time. Alternatively one can use historical record in order to obtain an estimation of completion rate. However because historical records are used, the indicator is mainly a result of past events rather than current events.

Our method is entirely based on the population of drop-out students during an academic year x. This population can be determined when a regular enrollment period of academic year $x + 1$ has ended. In this way substantial delay is avoided and the most up-to-date information that is available is used to estimate completion rate. The method emphasis the drop-outs during the academic year x and therefore mainly reflects current information on school effectiveness. This is a major difference that distinguish the new proposed estimator from other indicators found in literature.

Moreover the estimator is independent of study duration and therefore particularly useful in post secondary institutions where it is more common that bachelor or master students take more time to finish their studies than stated in the standard curriculum.

Although the new estimator proposed offers new advantages on already existing estimators it is subject to limitations. Our method showed to meet the requirements of estimating completion rate only when some underlying assumptions were met which were discussed in section 3.3. Moreover the method stays subject to data limitation as is the case for cohort based estimators.

In section 4 a statistical framework is presented in which completion rates can be studied. We re-encountered the assumptions underlying the method in formula (7). Finally a method is proposed to test hypotheses concerning the size of bias when one is comparing a cohort study with the method developed in this paper.

Acknowledgement

We gratefully acknowledge our anonymous reviewers for their helpful comments and suggestions.

References

Astin, A. W. (2005). Making sense out of degree completion rates. *Journal of College Student Retention*, *7*, 5–17.

Boden, G. (2011-2012). Retention and graduation rates: Insights from and extended logitudinal view. *Journal of College Student Retention: Research, Theory and Practice*, *13*, 179–203.

Cox, D. (2006). *Principles of Statistical Inference*. Cambridge: Cambridge University Press.

Ensminger, M., & Slucarcick, A. (1992). Paths to high school graduation or dropout: A longitudinal study of a first-grade cohort. *Sociology of Education*, *65*, 95–113.

Greene, J. P., & Winters, M. A. (2002). *Public School Graduation Rates in the United States*. Technical Report 31 Manhattan Institute for Policy Research.

Haney, W. (2001). Revisiting the myth of the texas miracle in education: Lessons about dropout research and dropout prevention. Paper prepared for the Dropout Research: Accurate Counts and Positive Interventions Conference Sponsored by Achieve and the Harvard Civil Rights Project, Cambridge MA.

Hauser, R. M., & Koenig, J. A. (2011). *High School Dropout, Graduation, and Completion Rates: Better Data, Better Measures, Better Decisions*. Washington, D.C.: The National Academies Press.

Miao, J., & Haney, W. (2004). High school graduation rates: Alternative methods and implications. *Education Policy Analysis Archives*, *12*.

NLCB (2002). No Child Left Behind (NCLB) Act of 2001. *Pub. L. No. 107-110*, *115 Stat. 1425*.

Sauro, J., & Lewis, J. (2005). Estimating completion rates from small samples using binomial confidence intervals: comparisons and recommendations. In *Proceedings of the Human Factors and Ergonomics Society 49th annual meeting*.

Seastrom, M. M., Chapman, C., Stillwell, R., Daniel, M., Peltola, P., Dinkes, R., & Xu, Z. (2006). *User's Guide to Computing High School Graduation Rates, Volume 1: Review of Current and Proposed Graduation Indicators.*. Technical Report 604 National Center for Education Statistics.

Swanson, C. B., & Chaplin, D. (2003). *Counting High School Graduates when Graduates count: Measuring Graduation Rates under the High Stakes of NCLB*. Technical Report Education Policy Center, The Urban Institute.